Variability and Accuracy in Mammographic Interpretation Using the American College of Radiology Breast Imaging Reporting and Data System

Karla Kerlikowske, Deborah Grady, John Barclay, Steven D. Frankel, Steven H. Ominsky, Edward A. Sickles, Virginia Ernster

Background: **Several studies, which were limited by their small sample size and selection of difficult cases for review, have reported substantial variability among radiologists in interpretation of mammographic examinations. We have determined, in the largest study to date, intraobserver and interobserver agreement in interpreting screening mammography and accuracy of mammography by use of the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS).** ***Methods:*** **The mammographic examinations were randomly selected on the basis of original mammographic interpretation and cancer outcome from 71 713 screening examinations performed by the Mobile Mammography Screening Program of the University of California, San Francisco, during the period from April 1985 through February 1995. The final sample included 786 abnormal examinations with no cancer detected, 267 abnormal examinations with cancer detected, and 1563 normal examinations. Films were read separately by two radiologists according to BI-RADS. Cancer status was determined by contacting women's physicians and by linkage to the regional Surveillance, Epidemiology, and End Results Program.** ***Results:*** **There was moderate agreement between radiologists in reporting the presence of a finding when cancer was present ($\kappa = 0.54$) and substantial agreement when cancer was not present ($\kappa = 0.62$). Agreement was moderate in assigning one of the five assessment categories but was statistically significantly lower when cancer was present relative to when cancer was not present ($\kappa = 0.46$ versus 0.56; two-sided *P* = .02). Agreement for reporting the presence of a finding and mammographic assessment was twofold more likely for examinations with less dense breasts. Agreement was higher on repeat readings by the same radiologists than between radiologists. The sensitivity of mammography was lower with BI-RADS than with the original system for mammographic interpretation, but the positive predictive value of mammography was higher.** ***Conclusion:*** **Considerable variability in interpretation of mammographic examinations exists; this variability and the accuracy of mammography are neither improved nor diminished with use of BI-RADS. [J Natl Cancer Inst 1998;90: 1801–9]**

Several studies *(1–6)* have reported substantial variability among radiologists in interpretations of mammographic examinations and recommendations for management of breast lesions. Many of these studies are limited by the small number of screening examinations reviewed, the small number of cancer cases included, and the selection of difficult cases for review. With the implementation of mass screening for breast cancer, interest in improving the consistency and accuracy of mammographic interpretations is increasing.

The American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) *(7)* was designed to standardize the interpretation of mammographic examinations and the reporting of results by providing six well-defined assessment categories, four categories to describe breast density, 12 types of breast lesions, multiple standard descriptors of the morphology of breast lesions, and standard recommendations for follow-up. One study *(2)* determined the interobserver agreement for choosing terms to describe masses and calcifications using BI-RADS and found moderate agreement ($\kappa = 0.5$) among five radiologists on 60 abnormal mammographic examinations. Interobserver agreement for classifying breast lesions into one of five assessment categories (''benign,'' ''likely benign,'' ''intermediate,'' ''likely malignant,'' and ''malignant'') was fair to moderate ($\kappa = 0.4$) *(2).*

The reproducibility of BI-RADS mammographic assessment and breast density categories, morphologic descriptors of breast lesions, and recommendations for follow-up screening and for diagnostic tests has not been studied in a large sample of screening mammographic examinations that includes normal and abnormal examinations and a large number of examinations that subsequently resulted in the detection of invasive cancer or ductal carcinoma *in situ.* Determining the reproducibility of BI-RADS is important because it is currently being widely used in an effort to improve the consistency in reporting of mammographic results.

We used BI-RADS to determine intraobserver and interobserver agreement in interpreting screening mammographic examinations, describing specific mammographic lesions, and recommending follow-up tests. We also compared the accuracy of mammography based on the use of BI-RADS with that based on the original system for mammographic interpretation.

Affiliations of authors: K. Kerlikowske, D. Grady (Department of Epidemiology and Biostatistics, Department of Medicine, and General Internal Medicine Section, Department of Veterans Affairs), J. Barclay, V. Ernster (Department of Epidemiology and Biostatistics), S. D. Frankel, S. H. Ominsky, E. A. Sickles (Department of Radiology), University of California, San Francisco.

Correspondence to: Karla Kerlikowske, M.D., San Francisco Veterans Affairs Medical Center, General Internal Medicine Section, 111A1, 4150 Clement St., San Francisco, CA 94121 (e-mail; kerliko@itsa.ucsf.edu).

See ''Notes'' following ''References.''

Subjects and Methods

Subjects

The Mobile Mammography Screening Program of the University of California, San Francisco, is a low-cost breast cancer screening program that offers mammography to women of diverse ethnic backgrounds (54% nonwhite) in six counties in Northern California. Our study sample included women aged 30 years and older who underwent screening mammography during the period from April 1985 through February 1995. Screening procedures have been described in detail *(8–10)*. In brief, mammography was performed in a mobile van staffed by three certified radiologic technologists. Each woman completed a brief risk profile for breast cancer, and two standard mammographic views per breast were obtained with an accredited dedicated mammography unit (Mamex DC or Instrumentarium Alpha III). The risk profile for breast cancer included questions about personal and family histories of breast cancer *(10)*. Women were considered to have a family history of breast cancer if they reported having at least one first-degree relative (mother, sister, or daughter) with breast cancer. Women with a history of mastectomy were excluded. The study was approved by the Committee on Human Research of the University of California, San Francisco.

Original System for Mammographic Assessment and Determination of Cancer Status

All films were initially reviewed by a staff radiologist, a breast imaging fellow, and a radiology resident in training, with the staff radiologist giving the final interpretation. The age and breast cancer risk status of women were available at the time of interpretation for radiologists who elected to review such information. Original mammographic assessments were reported as ''normal,'' ''additional evaluation needed,'' ''suspicious for malignancy (biopsy recommended),'' or ''malignant'' by radiologic criteria. For the selection of study mammographic examinations, the latter three categories were considered to be an abnormal mammographic result.

Clinical outcomes for all women with screening examinations that were interpreted as abnormal were determined by contacting the woman's personal physician and searching the pathology and radiology databases at the University of California, San Francisco. One month after an abnormal examination, referring physicians were sent a standardized request for information regarding diagnostic procedures performed to evaluate abnormal mammography and the clinical outcome. If physicians did not respond to the mailed request within 1 month, they were contacted by telephone. The monthly computer-generated request for information to physicians resulted in nearly complete follow-up of all abnormal examinations, with only 0.4% of women with abnormal examinations lost to follow-up *(9,11,12)*. In addition, records of all women who underwent screening mammography were linked by computer to the regional Surveillance, Epidemiology, and End Results Program,[1] which collects population-based cancer data from nine counties in Northern California, to determine cancer outcomes. Women were considered to have breast cancer if biopsy results or reports to the Surveillance, Epidemiology, and End Results Program showed any invasive carcinoma or ductal carcinoma *in situ.*

Selection of Mammographic Examinations

From a consecutive sample of 71 713 screening mammographic examinations performed during the study period, we randomly selected screening examinations, according to the year of the examination, from women with 1) true-positive examinations (abnormal mammographic examinations that were performed within 13 months before the date of breast biopsy with a diagnosis of breast cancer), 2) false-positive examinations (abnormal mammographic examinations that did not result in a diagnosis of breast cancer within 13 months), 3) true-negative examinations (normal mammographic examinations that did not result in a diagnosis of breast cancer within 13 months), and 4) false-negative examinations (normal mammographic examinations that were performed within 13 months before the date of a biopsy with a diagnosis of breast cancer). To collect sufficient data on specific mammographic findings (such as masses and calcifications), we oversampled women with abnormal examinations so that for each woman with a true-positive examination, we randomly selected two women with a false-positive examination of either breast. Only one woman with a normal examination of both breasts was selected per woman with cancer. Of the 427 cancer cases, only 29% of the screening examinations from the 427 women with cancer were unavailable for our study. As a result of the random selection of examinations and availability of films, the distribution of mammographic interpretations by breast was 30% false-positive examinations (n = 786), 10% true-positive examinations (n = 267), and 60% normal examinations (n = 1563). All false-negative examinations (n = 35) in the sample population were included, which was 2% of the normal breast examinations.

BI-RADS Terms for Mammographic Interpretation

The following BI-RADS terms *(7)* and specific categories for each term were used for mammographic interpretations (Fig. 1): 1) breast density (''entirely fatty,'' ''scattered fibroglandular tissue,'' ''heterogeneously dense,'' or ''extremely dense''), 2) breast findings (''mass,'' ''calcification,'' ''density,'' ''focal asymmetric density,'' ''multiple bilateral masses,'' ''multiple bilateral calcifications,'' ''multiple bilateral calcifications and masses,'' ''architectural distortion,'' and ''other findings''), 3) characteristics of breast masses (''density,'' ''size,'' ''shape,'' and ''margin'') and of calcifications (''type'' and ''distribution''), 4) nine locations of the breast finding (''upper,'' ''upper outer,'' ''outer,'' ''lower outer,'' ''lower,'' ''lower inner,'' ''inner,'' ''upper inner,'' and ''central or retroareolar'') and the depth of the finding (''anterior,'' ''middle,'' or ''posterior''), 5) mammographic assessments (''normal;'' ''normal with benign finding;'' ''incomplete, additional evaluation needed;'' ''suspicious abnormality;'' or ''highly suggestive of malignancy''), and 6) follow-up recommendations (''normal interval screening,'' ''additional mammographic views,'' ''ultrasound,'' ''fine-needle aspirate,'' or ''breast biopsy''). Of note, we used BI-RADS terminology with the following exceptions: 1) Mammographic assessment ''probably benign'' was not included as an assessment category because we only reviewed screening examinations; 2) mammographic assessment was not linked to recommendation to determine the distribution and frequency of first recommended follow-up tests; and 3) multiple bilateral masses, calcifications, or masses and calcifications were additional findings to the 12 included in BI-RADS *(7)*. Seven of the findings listed in BI-RADS were used only one time or not at all and thus were combined into one category, termed ''other findings.''

Radiologists' Experience and Study Preparation

The two participating radiologists were board certified and had 5 years and 12 years of experience, respectively, interpreting screening mammographic examinations; on average, they interpreted at least 1000 screening examinations per year. Radiologists used a computerized bar code and a software system designed for efficient data collection. Two training sessions were conducted before the start of the study. Before the first session, the radiologists reviewed BI-RADS terms and were trained to use the computer-based data entry system. Mammograms for 25 cases (50 breast mammographic interpretations), selected to represent a variety of breast findings and mammographic assessments, were read independently by the two study radiologists. After the first 50 breast examinations were read, we met to review the films, discuss disagreements between radiologists in film interpretation, and clarify definitions of mammography variables. A second group of 21 representative mammographic examinations were read by the two radiologists, and findings were again reviewed to resolve disagreements and to clarify definitions of BI-RADS terms.

Radiologists' Reading

For the study, screening examinations were read separately by the two radiologists who were blinded to the date of examination, to the original mammographic interpretation, and to the woman's age, risk profile for breast cancer, and cancer status. They were asked to assign a breast density for each woman (based on both breasts), to describe any breast findings, and to assign the mammographic assessment and the first recommended follow-up test separately for each breast. If more than one abnormality was noted in the same breast, radiologists were asked to describe each finding. They were asked to interpret examinations as they would in clinical practice with no time limitation for film interpretation.

Each radiologist read an average of 30 cases or 60 mammographic breast interpretations per session over a total of 48 sessions. Examinations were presented in random order with similar proportions of true-positive and false-positive examinations and true-negative and false-negative examinations per session. For measurement of intraobserver agreement, after the first six sessions, films were randomized again so that sessions 1 and 2 were reassigned randomly to sessions 7 or 8, sessions 3 and 4 were reassigned randomly to sessions 9 or 10, and sessions 5 and 6 were reassigned randomly to sessions 11 or 12. In addition, films were randomly reordered within each session. At least 6 weeks elapsed

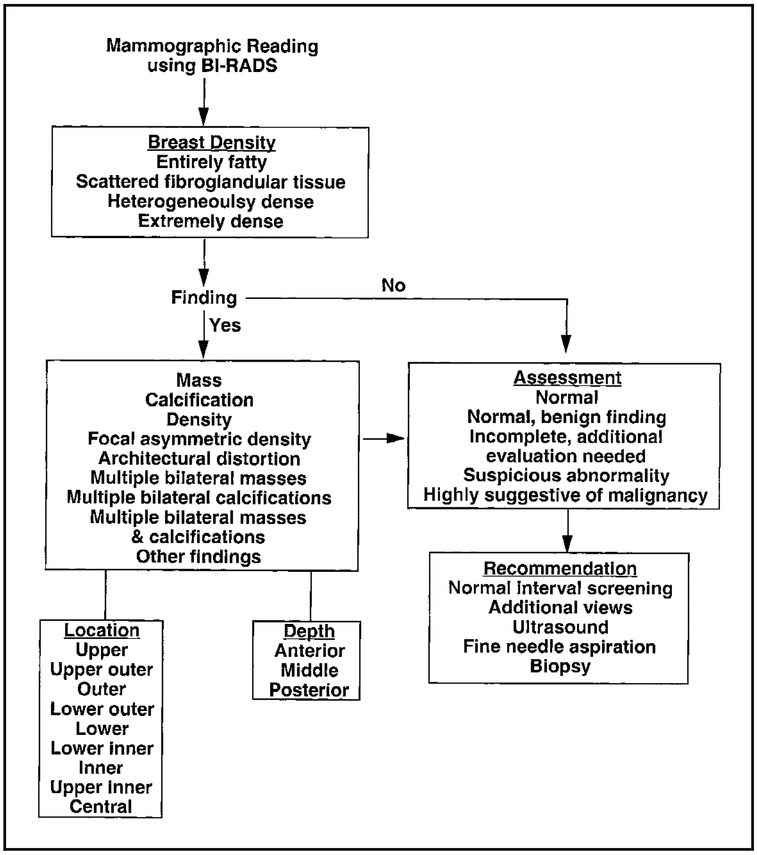

Fig. 1. Breast Imaging Reporting and Data System (BI-RADS) terms for mammographic interpretation.

before a screening examination was reread, as suggested by Metz *(13)*. Radiologists were unaware of the design of the study, the proportion of examinations that were originally interpreted as abnormal or represented cancer, or the fact that they were rereading mammographic examinations. The radiologists did not discuss their findings with each other during the study.

Data Analysis

We assessed intraobserver agreement based on the mammographic interpretation of 356 breasts by each radiologist in sessions 1 through 6 and again in sessions 7 through 12. Interobserver agreement was assessed on the basis of the readings by the two radiologists of 2616 breast examinations in sessions 1 through 6 and sessions 13 through 48. Variability was assessed for reporting the presence of a breast finding (finding or no finding) and the type of finding (nine categories) and for assigning breast density (four categories), mammographic assessment (five categories), and follow-up recommendation (five categories).

When a breast finding was noted, interobserver agreement was assessed for the location within the breast (nine locations), the depth of the finding (three categories), four specific characteristics of breast masses, and two specific characteristics of calcifications (*see* ‘‘BI-RADS Terms for Mammographic Interpretation’’ section *above* for a description of categories) *(7)*. For any pair of readings of the same breast, disagreement for an ordinal variable was considered to exist if there were any categorical differences between readings. The exception to this rule was for the location of breast findings where agreement was considered to exist if an observation was in the same area or an adjacent area; an observation in the central area was in agreement with all of the other eight locations. Agreement was assessed separately for mammographic interpretations of breasts with and without cancer.

The statistical test to assess the variability or the degree of agreement was the κ statistic *(14)*. Guidelines used for interpreting the amount of agreement were those reported by Landis and Koch *(15)*, where κ values were defined as follows: $\kappa = 0$ is poor agreement; $\kappa = 0.01–0.2$ is slight agreement; $\kappa = 0.21–0.40$ is fair agreement; $\kappa = 0.41–0.60$ is moderate agreement; $\kappa = 0.61–0.80$ is substantial agreement; and $\kappa = 0.81–1.00$ is almost perfect agreement. Similar scales for strength of agreement have been proposed by Brennan and Silman *(16)* and Fleiss *(17)*. A χ^2 test was used to determine equality between κ values *(17)*. Disagreement for mass size was measured by use of an intraclass correlation coefficient.

Logistic regression was performed to determine independent predictors of agreement. We calculated odds ratios (ORs) and 95% confidence intervals (CIs) for demographic and mammographic factors associated with the following three outcomes: 1) agreement in reporting the presence of a finding, 2) agreement in assigning a mammographic assessment, and 3) agreement in assigning a follow-up recommendation. For the logistic models, an average breast density reading was calculated for two breast readings by the same radiologist for the intraobserver models and for the two breast readings by different radiologists for the interobserver models. Entirely fatty breast density was assigned a value of 0, scattered fibroglandular tissue was assigned a value of 1, heterogeneously dense tissue was assigned a value of 2, and extremely dense tissue was assigned a value of 3. Mean breast density results were dichotomized so that low breast density scores ranged from 0 to 1 and high breast density scores ranged from 1.5 to 3. The average readings provide an independent measure of breast density and were used in the logistic model to measure the impact of breast density on agreement.

We compared the accuracy of using BI-RADS with the accuracy of using the original nomenclature for mammographic interpretation and the manner of reading mammography by calculating the sensitivity of mammography (number of true-positive examinations divided by the number of true-positive plus false-negative examinations) and the positive predictive value (PPV) of mammography (percent of women with abnormal screening examinations who were diagnosed with breast cancer) for each system. Because study radiologists read only 16.3% of the study films under the original nomenclature system, the number of cancers detected by them with this original system was too small ($n = 51$) to calculate the sensitivity and PPV of mammography for individual radiologists. Consequently, we pooled the results from seven radiologists (including the two study radiologists) who used the original system and calculated the sensitivity and PPV of mammography for the group based on the 208 cancers detected by first screening mammography, where comparison films were not available at the time of original interpretation. McNemar's test was used to compare the sensitivity of mammography when BI-RADS was used with the sensitivity of mammography when the original system of interpretation was used. The numerator and denominator for the PPV of mammography were not fixed because each radiologist could find a different number of examinations abnormal from the same study sample, which could lead to a different number of cancers detected. For this reason, the sample error of the difference between the PPVs of mammography using the two systems of interpretation was calculated from a bootstrap sample of 1000 replications of samples drawn with replacement and equal in size to the study population. All *P* values are two-sided.

Results

Primary Breast Findings, Breast Density, and Mammographic Assessment

The distribution of primary breast findings, breast density, mammographic assessments, and first follow-up recommendations is shown in Table 1. As determined by the two study radiologists using BI-RADS, among women with breast cancer, 27.8% and 41.8% were noted to have a mass, 38.6% and 39.8% were noted to have a calcification, and 0.4% and 13.7% were noted to have a focal asymmetric density. Compared with women with cancer, women without breast cancer who had a finding reported had a similar frequency of breast masses and focal asymmetric densities but had about half as many calcifications. The distribution of breast density did not vary by cancer status.

Among women with breast cancer, 21.8%–27.2% were interpreted as ‘‘normal’’ or ‘‘normal with a benign finding.’’ The most common first follow-up recommendation was additional mammography. Biopsy was recommended as the first test in 6.6%–14.1% of women with breast cancer. In this study, all

Table 1. Primary mammographic findings, assessment categories, and recommendations of two radiologists (radiologists [Rad.] A and B) among women with and without breast cancer

	% of total			
	Cancer* (n = 302)		No cancer (n = 2314)	
Parameter	Rad. A	Rad. B	Rad. A	Rad. B
Primary finding†				
Mass	27.8	41.8	26.4	37.6
Calcification	39.8	38.6	14.8	24.0
Focal asymmetric density	13.7	0.4	10.1	1.0
Other‡	18.7	19.3	48.8	37.3
Breast density				
Entirely fatty	4.0	2.7	5.6	3.7
Scattered fibroglandular tissue	48.0	43.1	42.4	42.1
Heterogeneously dense	38.0	51.0	41.4	48.8
Extremely dense	10.0	3.3	10.6	5.3
Assessment				
Normal	20.2	18.9	69.5	71.3
Normal with a benign finding	7.0	2.9	13.2	6.1
Incomplete, additional evaluation needed	51.7	53.4	17.1	22.4
Suspicious abnormality	13.6	13.7	0.2	0.3
Highly suggestive of malignancy	7.6	11.1	0	0
Recommendation				
Normal interval screening	27.2	21.9	82.7	77.2
Additional views	64.9	69.8	13.9	16.9
Ultrasound	1.3	3.3	3.5	5.0
Fine-needle aspiration	0	0	0	0
Biopsy	6.6	14.1	0	0

*Includes any invasive breast cancer or ductal carcinoma *in situ* detected within 13 months of abnormal or normal mammography.

†For the cancer group, n = 241 for Rad. A and n = 249 for Rad. B. For the no cancer group, n = 705 for Rad. A and n = 667 for Rad. B.

‡Other findings in the no cancer group were primarily bilateral masses, calcifications, or both.

women who had a biopsy recommended as her first follow-up test had breast cancer.

Overall for the study population, a greater proportion (71%–76%) of mammographic assessments were reported as ‘‘normal’’ or ‘‘normal with a benign finding’’ compared with the proportion originally assessed as ‘‘normal’’ (60%) by the original nomenclature and system of interpretation. In comparison, fewer false-positive examinations were reported for this study (15%–20%) compared with the proportion selected for the study sample (30%).

Intraobserver and Interobserver Agreement

Agreement between repeat readings by the same radiologist using BI-RADs is shown in Table 2. Overall, there was substantial agreement in reporting presence of a finding, describing the type of finding, and assigning a breast density and mammographic assessment, irrespective of cancer status. There was only moderate agreement in assigning a follow-up recommendation, and agreement was lower when cancer was present than when cancer was not present ($\kappa = 0.47$ versus 0.55; $P = .13$).

Agreement between two radiologists using BI-RADS in interpreting screening mammography is shown in Table 3. In general, agreement between radiologists was less than agreement on repeat readings by the same radiologist, especially if cancer was present. There was substantial agreement in reporting presence of a finding ($\kappa = 0.66$), and there was moderate agreement in assigning mammographic assessment ($\kappa = 0.58$) and first follow-up recommendation ($\kappa = 0.59$). There was only moderate agreement in reporting presence of a finding when cancer was present compared with substantial agreement when cancer was not present. In general, if a finding was present, there was moderate agreement for describing the type of primary finding, whether or not cancer was present. For describing type of primary finding, agreement between radiologists improved slightly ($\kappa = 0.63$) if describing the findings as a mass or as an asymmetric density were considered equivalent. There was also moderate to substantial agreement for describing breast density, irrespective of cancer status. When cancer was present, agreement in assigning mammographic assessment and selecting a first recommended follow-up test across five categories was moderate and lower than when cancer was not present (Table 3). Agreement for mammographic assessment across two categories (‘‘normal’’ or ‘‘normal with benign finding’’ versus ‘‘addi-

Table 2. Agreement between repeat readings by the same radiologist interpreting screening mammography films using the Breast Imaging Reporting and Data System

	All readings (n = 706)			Cancer (n = 72)			No cancer (n = 634)			
	Agreement			Agreement			Agreement			
Parameter	Observed, %	Expected, %*	κ (95% confidence interval)	Observed, %	Expected, %*	κ	Observed, %	Expected, %*	κ	*P*†
Finding/no finding	90	55	0.79 (0.74–0.83)	90	71	0.66	90	59	0.77	.20
Primary finding‡	78	25	0.71 (0.64–0.78)	87	42	0.78	75	23	0.68	.39
Breast density§	83	39	0.72 (0.66–0.78)	89	41	0.81	81	39	0.70	.11
Assessment category‖	86	49	0.73 (0.68–0.78)	75	33	0.62	87	55	0.72	.23
Recommendation¶	83	58	0.59 (0.56–0.63)	71	46	0.47	85	66	0.55	.13

*Expected agreement by chance alone.

†Comparison of κ statistic in cancer and no cancer group. All *P* values are two-sided.

‡The following nine findings were possible: mass, calcification, density, focal asymmetric density, multiple bilateral masses, multiple bilateral calcifications, multiple bilateral calcifications and masses, architectural distortion, and other findings (n = 61 cancer group; n = 193 no cancer group).

§The following four categories were possible: entirely fatty, scattered fibroglandular tissue, heterogeneously dense, and extremely dense (n = 323 for no cancer group).

‖The following five categories were possible: normal; normal with benign finding; incomplete, additional evaluation needed; suspicious abnormality; and highly suggestive of malignancy.

¶The following five categories were possible: normal interval screening, additional views, ultrasound, fine-needle aspiration, and biopsy.

Table 3. Agreement between two radiologists interpreting screening mammography films using the Breast Imaging Reporting and Data System

	All readings (n = 2578)			Cancer (n = 302)			No cancer (n = 2276)			
	Agreement		κ (95% confidence interval)	Agreement			Agreement			
Parameter	Observed, %	Expected, %*		Observed, %	Expected, %*	κ	Observed, %	Expected, %*	κ	*P*†
Finding/no finding	84	54	0.66 (0.63–0.69)	86	69	0.54	84	58	0.62	.21
Primary finding‡	66	23	0.56 (0.52–0.60)	71	30	0.58	64	22	0.54	.36
Breast density§	75	39	0.59 (0.55–0.62)	77	41	0.62	74	39	0.58	.39
Assessment category∥	78	48	0.58 (0.55–0.61)	65	34	0.46	80	54	0.56	.02
Recommendation¶	83	58	0.59 (0.56–0.63)	71	46	0.47	84	66	0.55	.13

*Expected agreement by chance alone.

†Comparison of κ statistic in cancer and no cancer group. All *P* values are two-sided.

‡The following nine findings were possible: mass, calcification, density, focal asymmetric density, multiple bilateral masses, multiple bilateral calcifications, multiple bilateral calcifications and masses, architectural distortion, and other findings (n = 222 cancer group; n = 500 no cancer group).

§The following four categories were possible: entirely fatty, scattered fibroglandular tissue, heterogeneously dense, and extremely dense (n = 302 cancer group; n = 1145 no cancer group).

∥The following five categories were possible: normal; normal with benign finding; incomplete, additional evaluation needed; suspicious abnormality; and highly suggestive of malignancy.

¶The following five categories were possible: normal interval screening, additional views, ultrasound, fine-needle aspiration, and biopsy.

tional evaluation needed; suspicious abnormality'' or ''highly suggestive of malignancy'') was substantial among those without cancer and moderate among those with cancer (κ = 0.61 versus 0.54; *P* = .25).

Agreement between radiologists using BI-RADS in describing location and characteristics of findings is shown in Table 4. If a finding was reported, there was moderate agreement in describing the location within the breast. When a mass was reported, there was substantial agreement in reporting mass depth but only fair to moderate agreement in reporting mass density, location, shape, and margin. Mass size within 2 mm was highly statistically correlated between radiologists (intraclass correlation coefficient = .82). Agreement for describing type and distribution of calcifications was fair to moderate.

Table 4. Agreement between radiologists interpreting screening mammography films using the Breast Imaging Reporting and Data System to describe location and characteristics of breast lesions

Parameter	No.	Observed, %	Expected, %*	κ (95% confidence interval
Location of primary finding†,‡	583	75	20	0.69 (0.64–0.73)
Mass				
Depth	244	85	39	0.76 (0.69–0.83)
Density	244	80	73	0.23 (0.04–0.41)
Location‡	244	58	23	0.46 (0.38–0.53)
Shape	244	60	34	0.40 (0.30–0.49)
Margin	244	69	27	0.58 (0.50–0.66)
Calcification				
Type	171	58	38	0.33 (0.21–0.46)
Distribution	171	78	60	0.46 (0.33–0.60)

*Expected agreement by chance alone.

†Inter-rater agreement for location of primary finding only including the following findings: mass, calcification, density, focal asymmetric density, architectural distortion, and other findings.

‡The following nine locations were possible: upper, upper outer, outer, lower outer, lower inner, inner, upper inner, central, or retroareolar.

Factors Associated With Intraobserver and Interobserver Agreement

Factors associated with intraobserver and interobserver agreement were similar when BI-RADS was used (Table 5). Agreement on reporting presence of a finding was more likely for mammographic examinations of less dense breasts (i.e., more fatty or radiolucent) than of more dense breasts (Table 5). This association was slightly stronger for repeat readings by one radiologist than for readings between radiologists. Presence of cancer did not influence agreement in reporting a finding for repeat readings by one radiologist or for readings between radiologists. There was slightly more agreement between radiologists in reporting a finding among women who reported a family history of breast cancer than among women who did not.

Agreement for mammographic assessment was twofold to threefold more likely in less dense breasts than in more dense breasts for repeat readings by one radiologist and for readings between radiologists (Table 5). Agreement on assessments across five categories was twofold greater when cancer was not present than when cancer was present for repeat readings by one radiologist and for repeat readings between radiologists; the association was slightly stronger for invasive cancer than for ductal carcinoma *in situ* (Table 5). When the five assessment categories were collapsed into two categories (''normal'' and ''normal with benign finding'' versus ''additional evaluation needed,'' ''suspicious abnormality,'' and ''highly suggestive of malignancy''), the presence of cancer no longer influenced agreement (data not shown). However, agreement was more likely in less dense breasts than in more dense breasts for repeat readings by one radiologist (OR = 2.3; 95% CI = 1.2–4.2) and for readings between radiologists (OR = 1.4; 95% CI =1.1–1.9). There was slightly greater agreement between radiologists in assigning mammographic assessment among women who reported a family history of breast cancer than among women who did not.

When cancer was not present, there was also greater agreement in which of the follow-up recommendations (five catego-

Table 5. Multivariable models of predictors of agreement between repeat readings for radiologists (intraobserver agreement, n = 706) and between radiologists (interobserver agreement, n = 2573)

Parameter	Odds ratio (95% confidence interval)	
	Intraobserver agreement	Interobserver agreement
Presence of finding*		
Breast density, low/high†	2.1 (1.2–3.7)	1.5 (1.2–1.9)
Age, per 10 y	0.8 (0.7–1.0)	1.0 (0.9–1.0)
DCIS, no/yes	1.7 (0.5–6.2)	0.9 (0.5–1.7)
Invasive cancer, no/yes	0.7 (0.3–2.2)	0.9 (0.6–1.3)
Family history, yes/no‡	1.4 (0.6–3.1)	1.4 (1.0–2.0)
Examination year, per y	1.0 (0.9–1.1)	1.0 (0.9–1.0)
Assessment§		
Breast density, low/high†	2.8 (1.7–4.7)	1.7 (1.4–2.2)
Age, per 10 y	0.8 (0.7–1.0)	0.9 (0.8–1.0)
DCIS, no/yes	1.8 (0.6–5.6)	1.8 (1.1–3.0)
Invasive cancer, no/yes	2.6 (1.3–5.1)	2.3 (1.7–3.0)
Family history, yes/no‡	1.3 (0.6–2.6)	1.4 (1.0–1.8)
Examination year, per y	1.0 (0.9–1.1)	1.0 (1.0–1.0)
Recommendation‖		
Breast density, low/high†	1.4 (0.8–2.3)	1.1 (0.9–1.4)
Age, per 10 y	0.9 (0.7–1.1)	1.0 (0.9–1.1)
DCIS, no/yes	2.0 (0.6–6.3)	2.4 (1.4–3.9)
Invasive cancer, no/yes	1.6 (0.7–3.5)	2.2 (1.6–3.0)
Family history, yes/no‡	1.2 (0.6–2.5)	1.6 (1.1–2.2)
Examination year, per y	1.0 (0.9–1.1)	1.0 (1.0–1.0)

*The following two categories were possible: finding present and finding absent. DCIS = ductal carcinoma *in situ.*

†Low breast density ranged from 0 to 1 and high breast density scores ranged from 1.5 to 3.

‡At least one first-degree relative (mother, sister, or daughter) with a history of breast cancer.

§The following five categories were possible: normal; normal with benign finding; incomplete, additional evaluation needed; suspicious abnormality; and highly suggestive of malignancy.

‖The following five categories were possible: normal interval screening, additional views, ultrasound, fine-needle aspiration, and biopsy.

ries) was selected for readings between radiologists but not for repeat readings by the same radiologist (Table 5). When agreement was assessed across two categories (normal interval follow-up versus any additional diagnostic test), the presence of cancer no longer influenced agreement, but agreement was more likely in less dense breasts than in more dense breasts for repeat readings by one radiologist (OR = 2.3; 95% CI = 1.2–4.2) and for readings between radiologists (OR = 1.4; 95% CI = 1.1–1.8). This suggests breast density is associated with initiating a diagnostic work-up to evaluate a breast abnormality and that cancer status is associated with the type of follow-up test recommended. Presence of a family history of breast cancer was associated with a greater agreement between radiologists in the follow-up recommendation selected.

Increasing age and examinations done in later years were not associated with agreement in reporting presence of a finding or with agreement in assessment or follow-up recommendation selected either for repeat readings by one radiologist or for readings between radiologists.

Accuracy of Mammographic Interpretation by Use of BI-RADS Versus the Original System

The sensitivity of mammography for each radiologist using BI-RADS was lower than the sensitivity for the group of radiologists using the original system for mammographic interpretation (Table 6), but the sensitivity based on the combined mammographic interpretations from both study radiologists (double reading) was similar to the sensitivity for the original system. The PPV of mammography for each radiologist using BI-RADS was higher than the PPV for the group of radiologists using the original system. Differences between systems were primarily in the ‘‘incomplete, additional evaluation needed’’ category where the PPV of mammography was higher for both radiologists using BI-RADS.

Discussion

We used BI-RADS to determine intraobserver and interobserver agreement in interpreting screening mammographic examinations, describing specific mammographic findings, and recommending follow-up screening or diagnostic tests. Mammography is used to screen women between the ages of 40 years and 69 years for a disease of low prevalence (annual incidence of invasive breast cancer is 15–34 cases of cancer per 10 000 women). Given this incidence, minimizing the variability in film interpretation should be a priority to maximize the accuracy of screening mammography and to minimize the number of false-positive evaluations. We found that having radiologists use BI-RADS for mammographic interpretation resulted in moderate agreement among radiologists in mammographic assessment

Table 6. Accuracy of mammography with the use of the Breast Imaging Reporting and Data System (BI-RADS) compared with the original system for mammographic interpretation

Parameter	BI-RADS			
	Radiologist A	Radiologist B	Radiologist A and B*	Original system, group reading†
Sensitivity, %	72.9	78.2	84.3	90.4‡
Overall PPV mammography, %§	15.8	13.4	12.6	9.5‡
PPV by abnormal interpretation, %§				
Incomplete, additional evaluation needed	11.8	9.7	8.7	5.5
Suspicious abnormality	77.7	67.1	64.2	46.7
Highly suggestive of malignancy	100	100	100	92.1

*Calculation of sensitivity and positive predictive value (PPV) assumes that all abnormal mammographic results by either radiologist would be further evaluated.

†Values for first screening mammography when radiologist does not have prior films for comparison.

‡All $P<.01$ when compared with either radiologist A or B.

§PPVs adjusted by sample weights for BI-RADS values.

and management recommendations and fair agreement in the use of descriptive terms to describe breast lesions. Agreement in mammographic interpretation was poorest among examinations from women with cancer and from those with mammographically dense breasts. The accuracy of mammography with the use of BI-RADS was similar to that of the original system for mammographic interpretation.

Under different study conditions and by use of various nomenclature systems for mammographic interpretation, other studies have also reported moderate to substantial variability in mammographic interpretation. Others have included only abnormal mammography or a higher percentage of abnormal mammography *(1,2,5,6),* had a higher proportion (38%–64%) of cases with cancer *(1,2,5,6),* and selected difficult cases, all of which would tend to result in more variability in mammographic interpretation *(1,2).* Although our study sample was enriched with abnormal and false-negative examinations, our sample more closely simulates the distribution of mammographic interpretations and cancer outcomes expected in practice than do previous studies. Moreover, the κ statistic, used to report degree of agreement, adjusts for the expected agreement that is influenced by the study sample distribution. Still, our results are comparable to those of others *(1,2,4),* showing that use of BI-RADS did not influence the consistency in mammographic interpretation. Specifically, using BI-RADS did not improve agreement in reporting presence of a finding or assigning a mammographic assessment or a recommendation for first follow-up tests compared with that reported by others.

As noted above, variability in mammographic assessment with the use of BI-RADS for reporting was greater among women with cancer than among women without cancer. There is better agreement among those without cancer because the majority of these women have normal mammographic assessments. Those with cancer may have findings suggestive of cancer that may be more or less obvious to radiologists. Depending on the radiologist's perception of the lesion and threshold for calling an examination abnormal, the screening examination may or may not be assessed as abnormal. Furthermore, even when radiologists perceive the same mammographic lesion among those with cancer, there may be disagreement in choice of assessment across the three possible categories of abnormal (''incomplete, additional evaluation needed,'' ''suspicious abnormality,'' and ''highly suggestive of malignancy''). Thus, there is more variability in assigning mammographic assessment among those with cancer because of differences between radiologists in perceiving a lesion as abnormal and because there are more choices for abnormal assessment categories (three) compared with normal assessment categories (two).

Whereas previous reports *(1,5)* have shown that 8%–10% of women without breast cancer were recommended for breast biopsy, in our study, all women who had a biopsy recommended as their first follow-up test had breast cancer. These findings could have occurred because radiologists in our study are more expert than in other studies and only recommended biopsies for lesions that were obviously cancer or because biopsy is usually reserved as a second or third procedure in evaluation of a mammographic abnormality.

Two prior studies have reported only fair to moderate agreement in describing morphologic characteristics of masses and calcifications; one study *(2)* used BI-RADS, and the other study *(18)* used descriptive terms agreed on by radiologists experienced with interpreting screening mammography. Our results are consistent with these findings and suggest that attempts to use BI-RADS terms to further describe and classify masses and calcifications may not be useful. The substantial variability in describing characteristics of masses and calcifications may be because the radiologists differed in their understanding of the definitions of BI-RADS terms or because the terms provided did not adequately describe the lesion, making the choice of descriptor difficult. Additional research is needed to better define lesions that are highly predictive of breast cancer and to determine whether training radiologists to specifically identify these lesions decreases variability in mammographic interpretation.

There is increasing interest in the effect of mammographic breast density on the risk of breast cancer and performance of mammography. Several studies *(19,20)* have reported that high mammographic breast density is an independent risk factor for breast cancer, and one study *(21)* reported a higher sensitivity of mammography among women aged 50 years and older who had primarily fatty (i.e., more radiolucent) breast density. Consistent with the findings of other reports *(4,22),* we found moderate agreement in classification of breast density across four categories. We also found that there was more agreement in reporting presence of a finding, assigning mammographic assessment, and initiating a diagnostic work-up among women with less dense breasts (i.e., more fatty or radiolucent). That is, there was more variability in film interpretation among women with more dense breasts. Because a greater proportion of younger women have dense breasts (80% for ages 40–49 years, 54% for ages 50–59 years, and 42% for ages 60–69 years) *(21),* there may be greater variability in film interpretation for these women.

To our knowledge, this is the first study to examine whether agreement in reading mammographic films varies with the year that the examination was performed. The quality of modern mammography has improved in the last decade *(23,24).* Thus, there could be a decrease in variability of mammographic readings as film quality improved. That is, improvement in film quality would result in improved resolution of breast images making breast lesions more apparent and thus easier to identify and describe. However, improvement in the film quality of mammography could result in an increase in the variability of mammographic readings with improved contrast, allowing for a greater number of breast lesions to be identified and described. We found that the year of the examination had no influence on agreement of reporting the presence of a breast lesion or in assigning a mammographic assessment or recommendation when accredited mammography units were used for screening. This suggests that additional incremental improvements in the technology of mammography may not greatly alter variability in mammographic interpretation.

The study radiologists were not aware of a woman's age or her family history of breast cancer before film interpretation because we wanted to evaluate the reproducibility of mammographic interpretation primarily based on the radiologist's perception of breast lesions and classification of lesions according to BI-RADS. On the one hand, if knowledge of clinical history alters a radiologist's level of diagnostic suspicion to report a breast lesion, not providing such information may have resulted in an underestimation of radiologists' agreement on film interpretation. On the other hand, not providing age and family history information may have improved agreement, since it has been shown that when knowledge of family history of breast cancer is available, radiologists tend to investigate more breast lesions without improving diagnostic accuracy *(25)*. We found that, by having radiologists unaware of a woman's age, their level of agreement in reporting the presence of a breast lesion or in assigning a mammographic assessment or recommendation was not influenced by a woman's age. Even though blinded to family history of breast cancer, there was slightly better agreement between radiologists in reporting presence of a breast lesion and in assigning a mammographic assessment or recommendation for women with a family history of breast cancer. These results suggest that breast lesions among women with a family history may be more apparent and easier to identify. In contrast, family history was not associated with agreement between repeat readings by the same radiologist, suggesting that the effect of family history may have been a chance result.

The accuracy of mammography was similar with the use of BI-RADS compared with that reported in the literature. Consistent with prior observer performance studies of the variability of mammographic interpretation *(1,5,24)*, we found that 25% of cancers were missed. Had all abnormal mammographic results by either radiologist been further evaluated, 16% of cancers would have been missed, or an additional 10% of the cancers would have been identified by double reading, similar to the rate reported by others *(26)*. When compared with the original system of film interpretation, the sensitivity of mammography was lower with the use of BI-RADS, but the PPV was higher. The lower sensitivity and higher PPV of screening mammography with the use of BI-RADS may have occurred because only one radiologist contributed to the mammographic interpretation compared with multiple reviewers during the original interpretation, because clinical history information was not available during the study and this information may have been helpful during the original interpretation, or because radiologists were less anxious about missing cancer in a study setting and used a higher threshold for calling an examination abnormal.

Our study has several strengths. First, our study population more closely represents women seen in clinical practice than other studies because screening examinations were randomly selected from a large screening population, had a greater proportion of normal examinations, and included both invasive cancer and ductal carcinoma *in situ*. Therefore, the level of variability for mammographic interpretation that we report may more closely approximate that expected in actual practice. Second, we linked our study population to the regional Surveillance, Epidemiology, and End Results Program to verify all cancer outcomes after normal and abnormal mammography results. The main limitation of our study is that only two radiologists participated in the study; therefore, our results may not be generalizable to all practicing general radiologists, in particular, those who read a low volume of screening examinations. Review of BI-RADS terms by study radiologists before the study may have contributed to the level of agreement that we report; therefore, our results also may not be applicable to radiologists less experienced with BI-RADS. Finally, had we measured intraobserver agreement at the end of the study, rather than at the beginning when radiologists had less experience with study procedures, we may have observed greater intraobserver agreement.

In summary, there was moderate agreement of mammographic interpretation between radiologists using BI-RADS. Variability in mammographic interpretation neither improved nor diminished with the use of a defined system (BI-RADS) for mammographic interpretation, and variability was unchanged over time despite technical improvements in mammography. Agreement in reporting presence of a breast finding and assigning mammographic assessment using the BI-RADS system is more likely for screening examinations from women with less dense breasts, and agreement in mammographic assessment and recommendation using the BI-RADS system is more likely when cancer is not present. Finally, use of BI-RADS did not influence the accuracy of mammography. To reduce variability of mammographic interpretation and potentially improve the accuracy of mammography, we need either better educational tools to communicate BI-RADS terms to radiologists or development of more effective criteria for reporting mammographic findings and selecting assessment categories. The American College of Radiology recently released an updated edition of BI-RADS *(27)* that includes mammographic illustrations of breast findings. This teaching device may improve the understanding of radiologists as to how and when to use BI-RADS terms and warrants testing to determine whether its use will decrease variability in mammographic interpretation.

References

(1) Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331: 1493–9.

(2) Baker JA, Kornguth PJ, Floyd CE Jr. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. Am J Roentgenol 1996;166:773–8.

(3) Ciccone G, Vineis P, Frigerio A, Segnan N. Inter-observer and intraobserver variability of mammogram interpretation: a field study. Eur J Cancer 1992;28A:1054–8.

(4) Vineis P, Sinistrero G, Temporelli A, Azzoni L, Bigo A, Burke P, et al. Inter-observer variability in the interpretation of mammograms. Tumori 1988;74:275–9.

(5) Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by U.S. radiologists. Findings from a national sample. Arch Intern Med 1996;156:209–13.

(6) Feldman J, Smith RA, Giusti R, DeBuono B, Fulton JP, Scott HD. Peer review of mammography interpretations in a breast cancer screening program. Am J Public Health 1995;85:837–9.

(7) American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS). 2nd ed. Reston (VA): American College of Radiology; 1995.

(8) Sickles EA, Weber WN, Galvin HB, Ominsky SH, Sollitto RA. Mammographic screening: how to operate successfully at low cost. Radiology 1986;160:95–7.

(9) Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993;270:2444–50.

(10) Sickles EA. The use of computers in mammography screening. Radiol Clin North Am 1987;25:1015–30.

(11) Sickles EA, Ominsky SH, Sollitto RA, Galvin HB, Monticciolo DL. Medical audit of a rapid-throughput mammography screening practice: methodology and results of 27,114 examinations. Radiology 1990;175:323–7.

(12) Monticciolo DL, Sickles EA. Computerized follow-up of abnormalities detected at mammography screening. Am J Roentgenol 1990;155:751–3.

(13) Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–33.

(14) Seigel DG, Podgor MJ, Remaley NA. Acceptable values of kappa for comparison of two groups. Am J Epidemiol 1992;135:571–8.

(15) Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

(16) Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992;304:1491–4.

(17) Fleiss J. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons; 1981. p. 218–22.

(18) Simpson W, Neilson F, Kelly PJ. The Northern Region Breast Screening Radiology Audit Group. Descriptive terms for mammographic abnormalities: observer variation in application. Clin Radiol 1995;51:709–13.

(19) Byrne C, Schairer C, Wolfe J, Parekh N, Salane M, Brinton LA, et al. Mammographic features and breast cancer risk: effects with time, age, and menopause status. J Natl Cancer Inst 1995;87:1622–9.

(20) Boyd NF, Byng JW, Jong RA, Fishell EK, Little LE, Miller AB, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. J Natl Cancer Inst 1995;87:670–5.

(21) Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Effect of age, breast density, and family history on the sensitivity of first screening mammography. JAMA 1996;276:33–8.

(22) Jong R, Fishell E, Little L, Lockwood G, Boyd NF. Mammographic signs of potential relevance to breast cancer risk: the agreement of radiologists' classification. Eur J Cancer Prev 1996;5:281–6.

(23) Conway BJ, McCrohan JL, Reuter FG, Suleiman OH. Mammography in the eighties. Radiology 1990;177:335–9.

(24) Conway BJ, Suleiman OH, Reuter FG, Antonsen RG, Slayton RJ. National survey of mammographic facilities in 1985, 1988, and 1992. Radiology 1994;191:323–30.

(25) Elmore JG, Wells CK, Howard DH, Feinstein AR. The impact of clinical history on mammographic interpretations. JAMA 1997;277:49–52.

(26) Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. Radiology 1994;191:214–44.

(27) American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS). 3rd ed. Reston (VA): American College of Radiology; 1998.

Notes

[1] *Editor's note:* SEER is a set of geographically defined, population-based central tumor registries in the United States, operated by local nonprofit organizations under contract to the National Cancer Institute (NCI). Each registry annually submits its cases to the NCI on a computer tape. These computer tapes are then edited by the NCI and made available for analysis.

Supported in part by Public Health Service (PHS) Breast Cancer Surveillance Consortium co-operative agreement 1U01CA63740 and PHS Breast Cancer SPORE grant P50CA58207 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services. Dr. Kerlikowske is supported, in part, by an American Cancer Society Career Development Award for Primary Care Physicians.

We thank Jocelyn Sperling for her exceptional efforts on this project.

Manuscript received March 20, 1998; revised June 29, 1998; accepted September 28, 1998.